Grant Messner
Section 2

Bi 1 HW 8

1a) For *E. Coli*, the genome is about 5 million base pairs in length, while the typical protein (gene) is about 300 amino acids (meaning 900 base pairs, since each amino acid contains 3 bases), giving us $\frac{5*10^6}{9*10^2} \approx \mathbf{6 * 10^3}$ **genes**. For *Drosophilia*, the genome is about 100 million base pairs in length, while the typical protein (gene) is still about 300 amino acids (900 base pairs), giving us $\frac{1*10^8}{9*10^2} \approx \mathbf{1 * 10^5}$ **genes**. The *E. Coli* gene estimate is pretty close to the given range of 4000 – 5500 genes. However, the *Drosophilia* gene count is not as close(it is a little more than half of the given gene count of 17000).

1b) We can use rectangles to approximate the perimeters (and therefore the lengths) of the introns:

| Intron number | Width (bp) | Height (bp) | Perimeter (bp) |
|---|---|---|---|
| 1 | 200 | 700 | 1800 |
| 2 | 50 | 50 | 200 |
| 3 | 100 | 200 | 600 |
| 4 | 50 | 100 | 300 |
| 5 | 200 | 600 | 1600 |
| 6 | 50 | 100 | 300 |
| 7 | 100 | 800 | 1800 |

This gives us a total intron length of (summing from perimeters in table) 6600 bp.

The exon lengths are simpler, as they are close to straight lines:

| Exon number | Length (bp) |
|---|---|
| 1 | 200 |
| 2 | 20 |
| 3 | 100 |
| 4 | 100 |
| 5 | 100 |
| 6 | 200 |
| 7 | 700 |

This gives us a total exon length of 1420 bp. Thus, the total length of the transcript is $1420 + 6600 = 8020$ pb. This means that our intron ratio is $\frac{6600}{8020} \approx .\mathbf{8}$, and our exon ratio is $\frac{1420}{8020} \approx .\mathbf{2}$.

1c) Since the spliced size is 758 aa, the prespliced size is (using the ratio from 1b) is $\frac{758\text{ aa}}{.2} \approx \mathbf{4000\ aa}$. This is equivalent to 12000 nt (since there are 3 nucleotides per amino acid). The nuclear division occurs every 8 minutes, but 4 of those minutes are spent in the M phase, leaving only 4 minutes for transcription (or, equivalently, 240 second). This gives us a minimum sampling rate of $\frac{12000\text{ nt}}{240\text{ sec}} = \mathbf{50}\ \mathbf{\frac{nt}{sec}}$.

2a) See Jupyter notebook

2b) Try $c(x) = Ae^{\sigma x}$

$$0 = D\frac{d^2c}{dx^2} - \frac{c}{\tau} = Ae^{\sigma x}(\sigma^2 D - \frac{1}{\tau})$$

$$\sigma^2 D - \frac{1}{\tau} = 0 \rightarrow \sigma = \pm\frac{1}{\sqrt{D\tau}}$$

$\boldsymbol{\sigma = -\frac{1}{\sqrt{D\tau}}}$ is the relevant solution, since $\sigma = \frac{1}{\sqrt{D\tau}}$ would imply that the concentration of the rightmost bin is the largest, when in reality it is zero in all of our graphs.

$A$ is simply the concentration of the leftmost bin, since $c(0) = Ae^{-\frac{1}{\sqrt{D\tau}}*0} = A$.

2c)

The French-Flag model states that:

$$c(x_{cf}) * G = c^*$$

Thus, $Ae^{-\frac{1}{\sqrt{D\tau}}x_{cf}}G = c^*$

We also know that $Ae^{-\frac{1}{\sqrt{D\tau}}x_{wild}} = c^*$, so we can solve this system and obtain:

$$Ae^{-\frac{1}{\sqrt{D\tau}}x_{cf}}G = Ae^{-\frac{1}{\sqrt{D\tau}}x_{wild}}$$

$$-\frac{1}{\sqrt{D\tau}}x_{cf} + \ln(G) = -\frac{1}{\sqrt{D\tau}}x_{wild}$$

$$\boldsymbol{x_{cf} = x_{wild} + \ln(G) * \sqrt{D\tau}}$$

3a - b) See Jupyter notebook

3c) There are a few outliers in the data, but with the same general trend as the theory curve. The blue outliers could have been caused by human error in clicking positions, and some of the red outliers could have been caused by some of the simplifications we made to our theory model (the French-Flag model).